German Record Linkage Center
WORKING PAPER SERIES
NO. WP-GRLC-2013-03 | JUNE 15, 2013

Linking Surveys and Administrative Data

Rainer Schnell

Linking Surveys and Administrative Data

Rainer Schnell
Methodology Research Unit
University of Duisburg-Essen
15.6.2013

Introduction

Surveys can be linked to other surveys and to administrative data. For most social scientists, data referring to individual persons or organizations are of primary importance. Linking such data sets of individual units is called record linkage. Record linkage seeks to identify the same objects in two different data bases using a set of common identifiers or unique combinations of variables. Record linkage is sometimes confused with data fusion, in which data of different units are merged to generate synthetic data sets. Furthermore, record linkage is not to be mistaken as augmenting surveys with aggregate data. Record linkage of surveys and administrative data aims to link records of the same unit of analysis (usually persons, sometimes organizations, commercial enterprises or patent applications).

Advantages and Disadvantages of Administrative Data

Linking micro level data of the same unit across data sets offers many opportunities for research, especially if administrative data can be linked to survey data. Administrative data has unique advantages (Brackstone 1987, Judson 2005, Judson 2007, Lane 2010). For example, administrative data usually covers nearly the complete population, providing large number of cases even for rare populations. The same aspect also allows the computation of small area statistics, which can be used to stabilize estimates from surveys (Lehtonen/Veijanen 2009). Depending on the kind of administrative data, the data quality may be higher (or lower) than those of survey data. If access to administrative data is granted, the linkage is inexpensive and fast. Given preprocessed data and established procedures, linkage can be done even for census operations within days. Administrative data does not generate additional respondent burden, it is not derogated by memory errors and rarely suffers from unit nonresponse. However, the use of administrative data has disadvantages (Judson/Popoff 2005). The most serious disadvantage is the limitation of the number of available variables. Administrative data bases usually don't contain subjective variables like attitudes. Furthermore, administrative data is based on administrative concepts, which may differ from research concepts (household definitions are an example). Administrative data bases have a time lag of months, sometimes years. Finally, even administrative personal identification numbers may have errors, so linkage might be far from perfect. For example, Judson (2005:440) reports error rates in social security numbers between 5 and 10%. So, linking records, in practice, is far from trivial.

Examples for Linking Surveys and Administrative Data

The amount of already existing information for record linkage is often surprising. However, to become useful for research purposes, it is not unusual that data from many different data bases must be extracted, transformed and very often linked across different data holders. Since the use of different data bases is not common in social science research, some examples of the use of record linkage with surveys may be helpful.

Generating Sampling Frames

In modern societies, most members of the population are listed in many different administrative data bases. Using the data bases for the determination of common subsets, record linkage can be used to construct sampling frames for special populations. For example, the research institute of the German Social Security Administration planned a panel study of the low income population. To study the mechanisms of entering and leaving the welfare population, a sampling frame of people with a high risk of entering the welfare population was needed. The persons had to be sampled before they entered the welfare population. Schnell (2007) suggested a series of record linkage operations, using municipal registries, welfare benefit registries and commercial credit risk data bases for sampling frame construction. A simplified version of this design has been used for the resulting panel study PASS, now one of the largest German population surveys.

Undercoverage Estimation for Census Operations

Harper/Mayhew (2012) tried to estimate the census undercoverage for a densely populated inner London borough. By using locally available administrative data (GP Register, School Census, Electoral Register, Council Tax Register, Council Tax and Housing Benefits, Births, Deaths, Housing Waiting List, Local Land and Property Gazetteer) they estimated a census undercoverage of 1.7%, with the undercoverage rate varying between age groups.

De-Duplication of Samples and Sampling Frames

The Los Angeles Women's Health Risk Study was a survey among street prostitutes in Los Angeles in the early 90s. One of the goals of the survey was the estimation of the prevalence of HIV infections in this population. The survey used an area-time-sample. The respondents received $25 for their participation. The 998 respondents were asked for self-generated identification codes based on names. Using elaborate record linkage techniques, Belin et al. (2004) estimated about 14.9% duplicates in the sample. However, they concluded that the main results of the study were not undermined by the presence of duplicates.

Constructing Panels Retrospectively

Linking subjects between independently collected data sets allows the construction of longitudinal data sets long after the data collection. A simple example is given by Jacobs/Boulis/Messikomer (2001). All physicians working in the US are represented in the AMA Physician Masterfile. Every year, one third of the records are updated with a survey. By using record linkage, Jacobs/Boulis/Messikomer (2001) built a longitudinal data file of the cohorts 1994 and 1998 with more than 500.000 physicians. This huge file permits the study of a rare event in a rare population: The probability of the change between different specialties within the medical profession.[1] Surprisingly, the probability of a professional change seems to be nearly independent of age.

Linking Panels over Time with Respondent Generated Codes

In many panel studies on sensitive topics, respondent-generated identification codes are used to link records across surveys because the use of identifiers like social security numbers or names is considered as privacy violation. Therefore, respondent-generated codes are based on stable but not obvious personal characteristics like respondents mothers' maiden names, the initials of close friends or the names of pets. Since these codes are error prone, usually a substantial number of cases are lost due to the codes. These losses may cause biased estimates. Schnell/Bachteler/Reiher (2010) suggested the use of record linkage for these kind of data: By using more components and linking the codes by the Levenshtein string distance function the losses could be reduced. The linking can

[1] Less than 0.4% of the employed Americans (about 129.7 million) are physicians; less than 1.1% of them reported a change. Finding them by sampling would have been a challenge.

be done with standard record linkage software. In two field experiments, the proposed procedure outperformed the methods previously applied.

Imputing Missing Survey Responses

Zanutto/Zaslavsky (2002) discussed the use of administrative data as a replacement for missing survey responses. Missing data can be replaced by data or the mean of multiple records from administrative records of the same case. Sometimes, the replacement may be useful only for a subset of cases. Finally, a statistical model based on administrative data might be used for the imputation of missing survey responses.

Validating Responses

An obvious application of record linkage is the validation of responses. Due to the large number of possible sources of response error (Weisberg 2005), the validity of survey responses is often doubtful. For example, in the practice of survey research, responses to questions on socially undesirable traits or behaviors are often considered more valid when the results show higher prevalences of undesirable behaviors. This assumption is rarely tested. One exception is the record linkage study of Maxfield/Weiler/Widom (2000). They compared the responses of 1196 young adults (mean 28.7 years) with administrative records on imprisonments. They conclude that 21% of all subjects with no history of arrest reported at least one arrest to the interviewers.

In a recent study Averdijk/Elffers (2012) linked a victimization survey of the city of Amsterdam to the police registration of crimes (n=8.887). Despite the widespread use of victimization surveys, such validation studies in criminology have been rarely reported during the last 25 years. 48% of the victimizations recorded by the police were not reported in the survey. 65% of the reported victimizations in the survey could not be found in the police records. For 18% of all respondents, survey and administrative records do not agree.

Checking Anonymity for Scientific Use Files

The demand for social research micro data has increased the number of available Scientific Use Files (SUF). The release of such files depends largely on the degree of anonymity which can be guaranteed to the respondents. So methods for an empirical disclosure risk assessment are needed. An obvious candidate for such a method is record linkage. Using the percentage of correctly linked record pairs between a SUF and a public available data set as measure for disclosure risk, Domingo-Ferrer/Torra (2003) demonstrate that record linkage for re-identification by cluster analysis of highly correlated variables do not require necessarily shared variables between the public available data set and a SUF. A similar approach has been published slightly earlier by Bacher/Brand/Bender (2002) for a German survey.

Technical Problems and Solutions

From a technical point of view, linking with a universally available unique personal identification number (PIDs) is ideal. In countries with universal available national PIDs (for example in Europe: Belgium, Denmark, Finland, Norway, Sweden) they cover the whole population. Under such conditions, linking different data bases is technically trivial. But these conditions are given only for very few countries. In practice, in most countries and for most linkage operations other identifiers have to be used. Most commonly, these are personal identifiers like name, date of birth or address. These identifiers have many disadvantages. They are not unique, therefore they must be used in combination. They are not stable, since persons change their names, place of residence and sometimes even their sex. Finally, many identifiers are recorded with errors. If the linkage is done on identical identifiers only, many actual matches will be missed. For example, Winkler (2009:362) reports that 25% of true matches in a census operation would have been missed by exact matching. In one region, 25% of the first names and 15% of the last names did not match perfectly.

Linkage with Imperfect Identifiers

If linkage has to be done with imperfect identifiers like names, things get a bit more complicated. In a first step, the identifiers have to be standardized. This step includes the conversion to the same codetable used for representing characters, removing titles and punctuation, transforming to uppercase, removing umlauts, replacing nicknames and so on. This preprocessing step can not be done automatically and is therefore always more tedious and lasts longer than expected by beginners.

Then a measure of similarity between names (strings) is needed. There are many different string similarity measures, but in many record linkage studies, the Jaro-Winkler (Winkler 1995) string similarity measure has shown a superior performance. For example, Schnell/Bachteler/Bender (2003) reported an increase of correctly linked pairs from 3.0% with exact matching to 32.1% with Jaro-Winkler.

Then the measures of similarity of different identifiers have to be combined to make the decision if a potential pair should be considered a match. Even today social scientists sometimes simply sum up the similarity scores and decide ad hoc on a similarity threshold. This is a suboptimal procedure since different identifiers should have different weights in deciding the match status. There are formal methods for determining the optimal weights of identifiers in this decision. These methods are based on a statistical decision theory for matching, described by Fellegi/Sunter (1969). The techniques for estimating the weights are now quite elaborated and beyond the scope of this paper. The application of these optimal weights for a decision on potential matching records is called "probabilistic record linkage" (see Herzog/Scheuren/Winkler 2007 for a textbook). There are many programs available for estimating the optimal parameters and performing probabilistic record linkage (details can be found at the end of the paper). However, probabilistic record linkage is far from being an automatic procedure, since preprocessing is usually laborious, estimating optimal parameters needs a lot of experimentation and after the linkage a clerical processing of unresolved pairs is needed. In practice, probabilistic record linkage is usually an iterative process. Depending on the quality of the identifiers, the size of data sets and the costs of producing false positive links (linking records, which do not belong together) and false negatives (missing true links), the process may take many months of labor.

Privacy Preserving Record Linkage

If record linkage needs to be done with personal identifiers, privacy concerns increase. Therefore, record linkage with encrypted identifiers is widely used, for example in medical research contexts like cancer registries. If record linkage is performed without revealing any information which can be used to identify the persons whose data is linked, it is called privacy preserving record linkage (PPRL) or private record linkage. The most simple variants of PPRL encrypt the identifiers, then linkage is done with the encrypted identifiers. This technique is highly secure, but even one different character in a name will result in two very different encrypted identifiers. Therefore, linking encrypted identifiers with exact matching will miss many true links between two files. Since the probability of slight changes in identifiers might depend on variables of interest, the missed true links might be different from detected links. For example, names in migrant populations often require transliteration; rules for transliteration might vary between data base systems. Furthermore, slight changes in names can be used intentionally by persons not wishing to be linked across data bases. Under such conditions, the successfully linked pairs are not a random sample of all pairs.

Because of this problems, it is common practice to replace string identifiers by an algorithmic substitution with special pseudonyms before encryption. These special pseudonyms are called phonetic codes. The most widely used phonetic code in the English speaking countries is the nearly 95 years old "Soundex" (see for example Christen 2012:74). A phonetic code maps similarly pronounced strings to the same key. For example, the names Engel, Engall, Engehl, Ehngel, Ehngehl, Enngeel, Ehnkiehl and Ehenekiehl all produce the same code (in this example: E524).

For privacy preserving record linkage, these codes are then encrypted with a cryptographic function (Borst/Allaert/Quantin 2001). There are many variants of this simple technique, resulting from different choices of preprocessing of strings, phonetic codes and cryptographic functions. However, the main problems of this approach are the same for all variants: These codes do not allow a similarity computation of the string identifiers of a potential pair: Either the code matches or not. Therefore, this approach suffers from missing many potential matches. Finally, within a group of potential pairs with the same code such as the example code E524 mentioned above, cases with identical other identifiers (birthday etc.) can not be separated. Under such conditions, this approach will produce false positive links. Many procedures for privacy preserving record linkages have been suggested (for an overview: Christen 2012:199-207). In real world settings only a few of these procedures can be used. A procedure which has been used successfully in different practical settings will be described below.

Privacy Preserving Record Linkage with Cryptographic Bloom Filters. Schnell/Bachteler/Reiher (2009) suggested a new method for the calculation of similarity between two encrypted strings for the use in record linkage procedures (Safelink). This method is based on the idea of splitting an identifier into substrings and mapping the set of substrings to a binary vector. Only these vectors are used for linkage. Since a one-way mapping is used, persons can not be identified by the vectors. The remaining of this subsection will give some technical details.

In general, the substrings of a string consisting of subsequent letters are called "n-grams". Usually, a string is extended on both ends with blanks before splitting into n-grams ("padding"). Padding is useful to distinguish n-grams in the middle of a name from n-grams at the beginning and end of a name.

For example, the set of 2-grams (called bigrams) of the name SMITH is the set _S,SM,MI,IT,TH,H_. This set of bigrams is mapped with a function to a vector (see figure 1). In this procedure, the functions are so called cryptographic hash functions (HMACs). One HMAC named MD5 is known to many computer users today, since it is also used as a checksum algorithm for CD-ROMs. HMACs are one of the building blocks in modern cryptography, a description can be found in any modern textbook on cryptography (for example, Stallings 2011). For record linkage, a variant of HMACs with a password is used ("keyed HMACs").

HMACs are one-way functions, so that two different inputs will be mapped to two different outcomes, but there is no way finding the input given only the outcome. Schnell/Bachteler/Reiher (2009) proposed mapping each n-gram with several different HMACs. The result of the function is mapped to a long vector of bits, initially all set to zero. The combination of a bit-vector with hash functions is called Bloom filters in computer science (Bloom 1970 invented this combination for an entirely different purpose).

Encoding the n-grams of a string with many HMACs to a Bloom filter has many advantages. The most important advantage is the fact, that given the Bloom filters alone, the initial string can not be reconstructed. Since only the Bloom filters are used for linkage, the identifiers are encrypted. But this mapping allows the computation of the similarity of two strings by using the Bloom filters only.

The procedure is best explained using an example. If we want to compare the similarity of the names "Smyth" and "Smith", we can use a standard string similarity measure like the Dice-coefficient. The Dice-similarity of two unencypted strings can be determined as

$$D_{a,b} = \frac{2h}{(|a| + |b|)},$$

where h is the number of shared bigrams and $|a|$, $|b|$ is the number of n-grams in the strings a, b. For example, the bigram similarity of "Smith" and "Smyth" can be computed by splitting the names into 2 sets of 6 bigrams each ({ $\_s, sm, mi, it, th, h\_$ } and $\{\_s, sm, my, yt, th, h\_\}$), counting the shared bigrams $\{\_s, sm, th, h\_\}$ and computing the Dice-coefficient as $\frac{2\times4}{(6+6)} \approx 0.67$.

If we want to compare the two strings with a Bloom filter encoding, we could for example use

bigrams and Bloom filters with 30 bits and two HMACs only (in practice, Bloom filters with 500 or 1000 bits and 15 to 50 hash functions are used). Figure 1 shows the encoding for this example. In both Bloom filters 8 identical bits are set to 1. Overall 1+10 bits are set to 1. Using the Dice-Coefficient, the similarity of the two Bloom filters is computed as $2*8/(11+10) \approx .76$. The similarity of two Bloom filters for two totally different names like SMITH and BLACK is much closer to zero. In general, the similarity between two names can be approximated by using the Bloom filters only. Therefore, the encoding of strings to Bloom filters as proposed by Schnell/Bachteler/Reiher (2009) allow the computation of string similarity with encrypted identifiers. A privacy preserving record linkage can be done by using the Bloom filters for standard identifies like names and addresses. By also encrypting numerical identifiers like the date of birth as strings, a privacy preserving record linkage allowing errors is possible.

Privacy Preserving Record Linkage with a Cryptographic Long-term Key. In some legal environments, record linkage has to be done with exactly one identifier, for example a PID may be required by law. For such environments, Schnell/Bachteler/Reiher (2011) suggested the use of the encoding described above, but with one common Bloom filter (called a cryptographic long term key or CLK) for all identifiers. So first name, last name, addresses, sex, date of birth etc. are all mapped to the same Bloom filter. For the mapping of each identifier, a different password is applied. Furthermore, the number of hash-functions can be varied between identifiers, for example to reflect the assumed discriminating power of identifiers for possible pairs.

Simulations and real world applications show a remarkable performance of the CLK. For example, in an application of a German cancer registry (Richter 2013), a CLK based on first name, last name, date of birth, zip-code and place of residence was used to link two files with 138.142 and 198.475 records. Considering only exact matches on the CLK, 18 pairs were found which had been not matched before.

In general, record linkage based on separately encoded Bloom filters should perform slightly better than a linkage based on CLKs of the same identifiers. However, CLKs have an obvious advantage over separate identifiers: An attack on CLKs is much more difficult than an attack on Bloom filters. One successful attack on cryptographic Bloom filters has been reported in the literature (Kuzu et al. 2011, Kuzu et al. 2013). They reported the correct assignment of a few names to separate Bloom filters, given a correct random sample of identifiers. Even this partially successful attack is much harder with CLKs. Currently, CLKs seem to be the only modern Privacy Preserving Record Linkage technique which has been used in applied research settings.

Respondents Permission for Linkage

The legal framework for record linkage varies from country to country. However, it seems that in most countries privacy considerations could be avoided, if the permission of the respondents to link their survey data to administrative records could be obtained. However, depending on the survey conditions (topic, population, interviewers, kind of request, phrasing of the request), large differences in the proportion of consent to linking have been observed. For example, Sakshaug/Kreuter (2012) reported between 24% and 89% consent to linkage for different surveys. Since consenting to linkage requests can be seen as a special kind of nonresponse, this additional step in a survey response process has generated a recent deluge of studies on linkage consent (for examples, see Sala/Burton/Knies 2012, Sakshaug/Tutz/Kreuter 2013). Currently, the bias generated by non-consenting seems to be small, at least in Germany (Sakshaug/Kreuter 2012). However, it can be expected that consent rates will decrease with increasing number of requests; furthermore differences in effects due to populations, research topics and sponsors on consent rates and consent bias must be taken into account. So the effects of non-consenting persons on estimates must be evaluated for each project anew.

Legal Permissions for Linkage

If bias reduction is of primary importance, the use of privacy preserving record linkage tech-

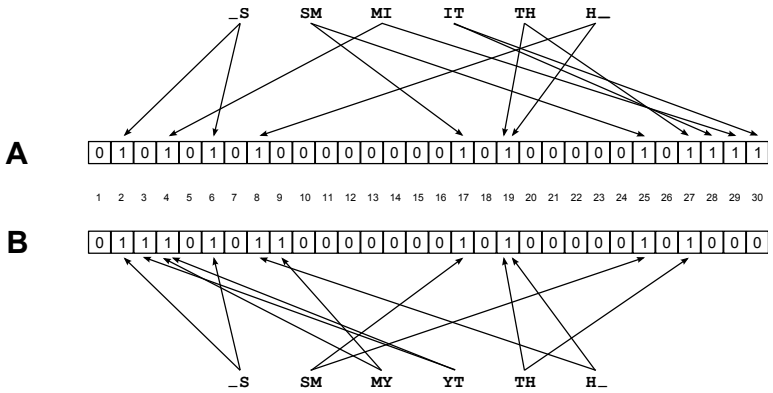

Figure 1. Example of the encoding of SMITH and SMYTH with two different cryptographic functions into 30-bit Bloom filters A and B (Schnell/Bachteler/Reiher 2009)

niques might be useful. In some countries, the legal framework allows data protection officers granting the linkage without consent, if public interests outweigh the privacy concerns. Arguing with privacy preserving linkage techniques can facilitate the negotiations with data protection officers. Demonstrating the factual anonymity with simulated data for subgroups with record linkage (see the paragraph above on checking anonymity) before the actual linkage has been done will also help. However, these discussions can take a long time. Informal inquiries with international research teams suggest that two years for discussions with data protection agencies and other stake holders is a reasonable estimate for national surveys.

Software

For serious work with large data files, standalone programs for record linkage should be used. During the last years, many commercial and open source solutions have become available. Since the requirements vary between research problems, a general recommendation can not be given. For evaluating record linkage software, Herzog/Scheuren/Winkler (2007, chapter 19) provide a checklist. A recent review of the current software has been given by Christen (2012, chapter 10). Only one of these programs can currently handle identifiers encrypted with Bloom filters: The so-called "Merge Toolbox" (MTB). This program (Schnell/Bachteler/Bender 2004) is available for free at the German Record Linkage Center (www.german-RLC.de).

Research Infrastructure for Record Linkage with Administrative Data

Recently, at least two European countries have established national research centers for linking administrative data bases. In the UK, the Administrative Data Liaison Service (ADLS) has been funded by the Economic and Social Research Council (ESRC) to support administrative data based research in the UK (see www.adls.ac.uk/about). The ADLS provides services and data access for the use of administrative data in the UK. In addition to these activities, the report of the Administrative Data Taskforce (2012) gave recommendations for fostering the use of administrative data for research. A very similar development in Germany was the establishment of Research Data Centers, initiated by the German Data Forum (www.ratswd.de). In addition, the German Research Foundation (DFG) funded the startup of an infrastructure for record linkage applications, the German Record Linkage Center (GRLC, for details see the homepage: www.german-rlc.de). The GRLC was established in 2011 to promote research on record linkage and to facilitate practical applications in Germany. The Center provides several services related to record linkage applications, for example acting as a data trustee for linkages. Within two years, more than 12 record linkage projects (including the large scale surveys SAVE, PASS, GSOEP and PAIRFAM) used the services of the GRLC.

Conclusion

Linking surveys and administrative data offers many unique advantages. Therefore, the linking of records of the same person or institution across different data bases will increase. The technical problems of linking even with imperfect identifiers have been solved, suitable software is widely available. The remaining problems are mainly caused by privacy objections. Very recent technical developments make record linkage with encrypted identifiers on micro data possible. The certification of this procedures by data protection agencies and their implementation in administrative contexts may take years. At least the same amount of time will be needed to include the linkage of administrative data to surveys into the standard toolbox of working social scientists. The establishment of national record linkage centers in the UK and in Germany are the first steps on this way.

References

Administrative Data Taskforce. (2012, December). *Improving access for research and policy.* ESRC Report. Swindon.

Averdijk, M., & Elffers, H. (2012). The discrepancy between survey-based victim accounts and police reports revisited. *International Review of Victimology*, *18*(2), 91-107.

Bacher, J., Brand, R., & Bender, S. (2002). Re-identifying register data by survey data using cluster analysis: an empirical study. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, *10*(5), 589-607.

Belin, T. R., Ishwaran, H., Duan, N., Berry, S. H., & Kanouse, D. E. (2004). Identifying likely duplicates by record linkage in a survey of prostitutes. In A. Gelman & X.-L. Meng (Eds.), *Applied bayesian modeling and causal inference from incomplete-data perspectives* (pp. 319–328). Hoboken: Wiley.

Bloom, B. H. (1970). Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, *13*(7), 422–426.

Borst, F., Allaert, F.-A., & Quantin, C. (2001). The Swiss solution for anonymous chaining patient files. In V. Patel, R. Rogers, & R. Haux (Eds.), *Proceedings of the 10th world congress on medical informatics* (pp. 1239–1241). Amsterdam: IOS Press.

Brackstone, G. J. (1987). Issues in the use of administrative records for statistical purposes. *Survey Methodology*, *13*(1), 29–43.

Christen, P. (2012). *Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection*. Berlin: Springer.

Domingo-Ferrer, J., & Torra, V. (2003). Disclosure risk assessment in statistical microdata protection via advanced record linkage. *Statistics and Computing*, *13*, 343–354.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*(328), 1183–1210.

Harper, G., & Mayhew, L. (2012). Using administrative data to count local populations. *Applied Spatial Analysis and Policy*, *5*(2), 97–122.

Herzog, T. N., Scheuren, F. J., & Winkler, W. E. (2007). *Data quality and record linkage techniques*. New York: Springer.

Jacobs, J. A., Boulis, A., & Messikomer, C. (2001). The movement of physicians between specialties. *Research in Social Stratification and Mobility*, *18*, 63–95.

Judson, D. H. (2005). Computerized record linkage and statistical matching. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (Vol. 1, p. 439-447). New York: Elsevier.

Judson, D. H. (2007). Information integration for constructing social statistics: History, theory and ideas towards a research programme. *Journal of the Royal Statistical Society: Series A*, *170*(2), 483–501.

Judson, D. H., & Popoff, C. L. (2005). Administrative records research. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (Vol. 1, p. 17-27). New York: Elsevier.

Kuzu, M., Kantarcioglu, M., Durham, E., & Malin, B. (2011). A constraint satisfaction cryptanalysis of bloom filters in private record linkage. In S. Fischer-Hübner & N. Hopper (Eds.), *The 11th privacy enhancing technologies symposium* (pp. 226–245). Berlin: Springer.

Kuzu, M., Kantarcioglu, M., Durham, E. A., Toth, C., & Malin, B. (2013). A practical approach to achieve private medical record linkage in light of public resources. *Journal of the American Medical Informatics Association*, *20*(2), 285–292.

Lane, J. (2010). Linking administrative and survey data. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (p. 659-680). Bingley: Emerald.

Lehtonen, R., & Veijanen, A. (2009). Design-based methods of estimation for domains and small areas.

In C. Rao (Ed.), *Handbook of statistics* (Vols. Volume 29, Part B, pp. 219–249). Amsterdam: Elsevier.

Maxfield, M. G., Weiler, B. L., & Widom, C. S. (2000). Comparing self-reports and official records of arrests. *Journal of Quantitative Criminology*, *16*(1), 87-110.

Richter, A. (2013, January). *Ergebnis des Abgleichs Nr. 2.* Memo, Cancer Registry Lübeck.

Sakshaug, J., Tutz, V., & Kreuter, F. (2013). Placement, wording, and interviewers: Identifying correlates of consent to link survey and administrative data. *Survey Research Methods*, *2*, in print.

Sakshaug, J. W., & Kreuter, F. (2012). Assessing the magnitude of non-consent biases in linked survey and administrative data. *Survey Research Methods*, *6*(2), 113–122.

Sala, E., Burton, J., & Knies, G. (2012). Correlates of obtaining informed consent to data linkage: Respondent, interview, and interviewer characteristics. *Sociological Methods & Research*, *41*(3), 414–439.

Schnell, R. (2007). Alternative Verfahren zur Stichprobengewinnung für ein Haushaltspanelsurvey mit Schwerpunkt im Niedrigeinkommens- und Transferleistungsbezug. In M. Promberger (Ed.), *Neue Daten für die Sozialstaatsforschung* (pp. 33–59). Nürnberg: Bundesagentur für Arbeit.

Schnell, R., Bachteler, T., & Bender, S. (2003). Record linkage using error-prone strings. In *Proceedings of the Section on Survey Research Methods, American Statistical Association* (pp. 3713–3717).

Schnell, R., Bachteler, T., & Bender, S. (2004). A toolbox for record linkage. *Austrian Journal of Statistics*, *33*(1–2), 125–133.

Schnell, R., Bachteler, T., & Reiher, J. (2009). Privacy-preserving record linkage using bloom filters. *BMC Medical Informatics and Decision Making*, *9*(41), 1–11.

Schnell, R., Bachteler, T., & Reiher, J. (2010). Improving the use of Self-Generated identification codes. *Evaluation Review*, *34*(5), 391-418.

Schnell, R., Bachteler, T., & Reiher, J. (2011). *A novel Error-Tolerant anonymous linking code* (Working Paper No. WP-GRLC-2011-02). Nuremberg: German Record Linkage Center.

Stallings, W. (2011). *Cryptography and network security: Principles and practice* (5th ed.). Boston: Prentice Hall.

Weisberg, H. F. (2005). *The total survey error approach: A guide to the new science of survey research*. Chicago: The University of Chicago Press.

Winkler, W. E. (1995). Matching and record linkage. In B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge, & P. S. Kott (Eds.), *Business survey methods* (pp. 355–384). New York: Wiley.

Winkler, W. E. (2009). Record linkage. In D. Pfeffermann & C. Rao (Eds.), *Handbook of statistics vol. 29a, sample surveys: Design, methods and applications* (pp. 351–380). Amsterdam: Elsevier, North-Holland.

Zanutto, E., & Zaslavsky, A. (2002). Using administrative records to impute for nonresponse. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 403–415). New York: Wiley.

IMPRINT

Publisher

German Record-Linkage Center
Regensburger Str. 104
D-90478 Nuremberg

Template layout

Christine Weidmann